

Submitted to *Genetics* February 2, 1995

Properties of Statistical Tests of the Neutral Model in Molecular Evolution

Katy L. Simonsen*, Gary A. Churchill^{*,†}, and Charles F. Aquadro[‡]

BU-1277-M

March 1995

*Center for Applied Math, [†]Biometrics Unit, [‡]Section of Genetics and Development,
Cornell University, Ithaca NY 14853

Corresponding Author:

Katy Simonsen
Center for Applied Math
Engineering and Theory Center
Cornell University
Ithaca NY 14853

E-mail: katy@cam.cornell.edu
Office: (607) 255-3399
Home: (607) 277-1788
Fax: (607) 255-9860

Abstract

A class of statistical tests of the neutral model of molecular evolution is studied to determine their size and power properties. The class includes Tajima's D statistic, as well as the D^* and F^* tests proposed by Fu and Li. A new method of constructing critical values for these tests is described. Simulations indicate that Tajima's test is generally most powerful against the alternative hypotheses of selective sweep, population bottleneck, and population subdivision, among tests within this class. However, even Tajima's test can detect a selective sweep or bottleneck only if it has occurred within a specific interval of time in the recent past, or population subdivision only when it has persisted for a very long time. For greatest power against the particular alternatives studied here, it is better to sequence more alleles than more sites.

INTRODUCTION

Given a set of aligned DNA sequences from a sample of n individuals of the same species, we would like to make inferences about the evolutionary history of the species. The neutral equilibrium model of sequence evolution is often considered as a null hypothesis against which specific alternative models can be compared. The neutral hypothesis is rejected if the observed data are unlikely to arise under this model. A problem of interest is to construct appropriate test statistics that will reject the neutral model with high probability when specific alternative models hold. We consider a class of test statistics that includes Tajima's D statistic (1989a) and the D^* and F^* tests proposed by Fu and Li (1993). The power properties of these tests against specific alternative hypotheses are studied using simulated data to determine how often and under which alternatives each test is able to reject the neutral model.

Critical values (rejection regions) of statistical tests are determined by the distribution of the statistics under the null hypothesis. The distributions of the test statistics we wish to examine are not known, but we can sample from these distributions by simulating data from the neutral model. Estimating the critical values is complicated because the distributions depend on the unknown value of a parameter θ which is proportional to the product of the effective population size and the mutation rate.

Our goal is to determine which statistical tests are most powerful against different alternatives and to determine the sample sizes necessary to achieve a reasonable power. We also address the issue of larger sample sizes versus greater number of sites sequenced with respect to improving statistical power.

This work was motivated in part by studies of natural populations of *Drosophila*, which have shown that levels of DNA polymorphism observed for a gene region are strongly correlated with regional rates of recombination, e.g., (AGUADÉ *et al.* 1989; STEPHAN and LANGLEY 1989; BEGUN and AQUADRO 1991; BERRY *et al.* 1991; BEGUN and AQUADRO 1992; AQUADRO *et al.* 1994). One hypothesis to explain this correlation is that hitchhiking associated with the fixation of advantageous mutations leads to a selective sweep and a resulting reduction of linked neutral variation, e.g. (KAPLAN *et al.* 1989). However, in most of these cases Tajima's D test did not reject the neutral model. This suggests the following

question: is Tajima's D powerful enough to detect selective sweep events? If Tajima's D often fails to reject the neutral model even after a selective sweep, then the selective sweep explanation remains viable for these data. If, however, a selective sweep event always results in a significant Tajima's D test, while a significant D is not observed in the data, it is unlikely that a selective sweep can completely explain the reduced levels of polymorphism. A second hypothesis is that reduced variation in regions of low recombination may result from the elimination of deleterious mutations, a process termed "background selection" by (CHARLESWORTH *et al.* 1993). This hypothesis also appears capable of predicting the observed correlation between variation and recombination, given sufficient latitude in the choice of evolutionary parameters, e.g., (WIEHE and STEPHAN 1993; AQUADRO *et al.* 1994; CHARLESWORTH 1994), (R. R. HUDSON and N. L. KAPLAN, personal communication). Background selection is not predicted to have an appreciable effect on the frequency distribution of standing variation, and hence on Tajima's D , since its effect is basically a reduction in the effective population size for gene regions with low recombination (CHARLESWORTH 1994). Thus, background selection is a possible explanation for the observed non-significant Tajima's D statistics.

Here, we first describe the coalescent model of neutral evolution. In the Methods section, we first describe a class of test statistics and a method by which critical values for statistical tests of the neutral model can be obtained; then, we describe how data can be simulated under alternatives to the neutral model. The Results section summarizes the outcome of these simulations, showing the effect of these alternatives on the distributions of the test statistics and their power. In the final section we discuss the implications of these results to performing statistical tests, and how they relate to the selective sweep hypothesis.

The Neutral Model

The neutral data were generated according to the coalescent model as described by (HUDSON 1990; HUDSON 1993). This model is based on the standard Wright-Fisher model, and makes the following assumptions:

1. a large constant diploid population size of N individuals or $2N$ alleles (where $N^2 \gg N$)
2. random mating
3. non-overlapping generations
4. no recombination
5. an infinite-sites, constant rate neutral mutation process: an offspring differs from its parent allele by a Poisson-distributed number of mutations with mean μ

Under these assumptions, the probability that two particular individuals have the same parent in the previous generation is $1/2N$. The probability that any two individuals in a sample of size j have the same parent is $p = \binom{j}{2} / 2N$. Thus, for a sample of j individuals in the current population, the probability that the first coalescent event between any two

of them occurs exactly $t + 1$ generations ago is $p(1 - p)^t$. That is, the time in generations during which there are exactly j lineages in the genealogy of the sample is geometrically distributed with mean $1/p$. It is convenient to treat time as a continuous random variable. To this end we approximate the geometric distribution with an exponential distribution with the same mean, since $p(1 - p)^t \approx pe^{-pt}$ for small p and large t . The assumption (1) that $N^2 \gg N$ ensures that p is sufficiently small. It is also convenient to measure time in units of $2N$ generations, with the result that p is replaced by $\binom{j}{2}$. Thus the time t_j in units of $2N$ generations during which there are exactly j lineages is exponentially distributed with mean $1/\binom{j}{2}$. The total time in the tree, T_{tot} , is equal to $\sum_{j=2}^n jt_j$.

The number of mutations that occur on a lineage of length t is, by assumption (5), Poisson-distributed with mean $2N\mu t = \theta t/2$, where $\theta = 4N\mu$. The assumption of infinite sites ensures that each mutation is observed as a polymorphic or segregating site. Therefore the number S of segregating sites in a sample is Poisson-distributed with mean $\theta T_{tot}/2$. However, as Hudson (1993) has pointed out, the fact that the true value of θ for data sets is unknown presents a problem when using simulation to estimate critical values for a test. Three methods of generating data are described in (HUDSON 1993): conditioning on θ , conditioning on θ and S , and conditioning on S . While the first method is the one consistent with our model, it cannot be used when θ is unknown. The other two methods require slightly different null and alternative hypotheses than we wished to use. Instead, we use the information contained in S to compute a range of values for θ that are consistent with the observed data. We then use values of θ in this interval to simulate the test statistic under the neutral model, and thus obtain critical values.

METHODS

Statistical Tests

From n nucleotide sequences, statistics such as S , the number of segregating sites, k , the average number of pairwise differences, and η_s , the number of singletons, may be calculated. These are random variables whose distribution depends on a parameter θ whose value is unknown, and each provides an unbiased estimate of θ . Let

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i}, \quad b_n = \sum_{i=1}^{n-1} \frac{1}{i^2}. \quad (1)$$

Under the neutral model, $E(S) = a_n\theta$, $E(k) = \theta$, and $E(\eta_s) = \left(\frac{n}{n-1}\right)\theta$. Their variances are

$$Var(S) = a_n\theta + b_n\theta^2 \quad (\text{WATTERSON 1975}) \quad (2)$$

$$Var(k) = \frac{(n+1)\theta}{3(n-1)} + \frac{2(n^2+n+3)\theta^2}{9n(n-1)} \quad (\text{TAJIMA 1983}) \quad (3)$$

$$Var(\eta_s) = \frac{n}{n-1}\theta + \left[\frac{2a_n}{n-1} - \frac{1}{(n-1)^2} \right] \theta^2 \quad (\text{FU and LI 1993}). \quad (4)$$

Therefore, S/a_n , k , and $\frac{n-1}{n}\eta_s$ are unbiased estimators of θ , and

$$m_1^2 = S(S-1)/(a_n^2 + b_n) \quad (5)$$

$$m_2^2 = \frac{3nk(3(n-1)k - n - 1)}{11n^2 - 7n + 6} \quad (6)$$

$$m_3^2 = \frac{(n-1)\eta_s(\eta_s-1)}{2a_n + n + 1} \quad (7)$$

are unbiased estimators of θ^2 .

In the following section we define a class of test statistics that includes three previously described test statistics and six new ones.

Test Statistics: From the three statistics S , k , and η_s , we can calculate test statistics such as those of Tajima (1989a):

$$D(k, S) = \frac{k - S/a_n}{\sqrt{u_T S + v_T S^2}} \quad (8)$$

and of Fu and Li (1993):

$$D^*(S, \eta_s) = \frac{S/a_n - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{D^*} S + v_{D^*} S^2}} \quad (9)$$

$$F^*(k, S, \eta_s) = \frac{k - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{F^*} S + v_{F^*} S^2}} \quad (10)$$

where the coefficients u and v are given in the Appendix. We refer to the coefficients for Tajima's statistic as u_T and v_T rather than u_D and v_D to distinguish them from those for Fu and Li's D statistic, which is not studied in this paper since it requires an outgroup. The formula for v_{F^*} given in the Appendix differs slightly from that given in Fu and Li (1993, unnumbered equation p. 702) due to a typographical error in their paper. Under the neutral model the three statistics D , D^* , and F^* all have expected value zero and variance approximately one.

Each of these statistics is constructed in the same way. Under neutrality, the three quantities S/a_n , k , and $\eta_s \left(\frac{n-1}{n}\right)$ all have expected value θ . Thus the difference between any two of these statistics will have expected value zero. The variances of the differences are of the form $\gamma\theta + \epsilon\theta^2$, where γ and ϵ depend on the statistic in question. The variance of the difference is estimated using S/a_n and m_1^2 as unbiased estimates for θ and θ^2 , respectively. The result is an estimated variance of the form $uS + vS^2$ where $u = \gamma/a_n - v$ and $v = \epsilon/(a_n^2 + b_n)$. Each test statistic is constructed by dividing the difference by the square root of its estimated variance.

The statistics D , D^* , and F^* use S and m_1^2 to estimate θ and θ^2 in the variance term $\gamma\theta + \epsilon\theta^2$. It is also possible to use k and m_2^2 or η_s and m_3^2 to make this estimate. S has been used because S/a_n has a smaller variance under neutrality than the other possibilities. In

non-neutral situations, however, the behavior of S , k , and η_s is more complex, so that k or η_s could make a better estimator of θ or θ^2 in some cases. We can construct six new test statistics, as follows.

$$T_2(k, S) = \frac{k - S/a_n}{\sqrt{u_{T_2}k + v_{T_2}k^2}} \quad (11)$$

$$T_3(k, S, \eta_s) = \frac{k - S/a_n}{\sqrt{u_{T_3}\eta_s + v_{T_3}\eta_s^2}} \quad (12)$$

$$D_2^*(k, S, \eta_s) = \frac{S/a_n - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{D_2^*}k + v_{D_2^*}k^2}} \quad (13)$$

$$D_3^*(S, \eta_s) = \frac{S/a_n - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{D_3^*}\eta_s + v_{D_3^*}\eta_s^2}} \quad (14)$$

$$F_2^*(k, \eta_s) = \frac{k - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{F_2^*}k + v_{F_2^*}k^2}} \quad (15)$$

$$F_3^*(k, \eta_s) = \frac{k - \eta_s \left(\frac{n-1}{n}\right)}{\sqrt{u_{F_3^*}\eta_s + v_{F_3^*}\eta_s^2}} \quad (16)$$

The subscript 2 or 3 indicates that the estimate of θ uses k and m_2 , or η_s and m_3 , respectively. The coefficients u and v are defined in the Appendix. The properties of these tests will be investigated along with those of the standard tests.

Hypothesis Testing Issues: Since neither S , k , nor η_s is a sufficient statistic for θ , the variance of any of the above test statistics will not be one and will vary with θ (HUDSON 1993). Thus, computed critical values for these test statistics must account for the unknown θ . Furthermore, even when θ is known, the exact distribution of the statistics under the null hypothesis is not. In order to perform two-sided tests of level α , the critical values required are the boundaries of a $(1 - \alpha)$ confidence interval for the test statistic. That is, for the statistic D we require D_U and D_L , independent of θ , such that the sum of the p-values $p_L = Pr_{H_0}(D \leq D_L)$ and $p_U = Pr_{H_0}(D \geq D_U)$ is less than or equal to α . Other authors have suggested methods to determine critical values as described below. We present an alternative method.

Tajima (1989a) computed critical values by assuming D to have a beta distribution with mean zero and variance one, scaled to the interval $[D_{min}, D_{max}]$. We note that since S and $k \binom{n}{2}$ are integers, D should follow a discrete, not a continuous distribution. Tajima's justification is based on a visual comparison between beta densities and histograms of simulated data. We have found that Tajima's critical values are often too conservative, particularly at the upper tail of the distribution. While it is true that the probability of false rejection

is not increased by using conservative critical values, it can result in a serious reduction in power. Thus this method of obtaining critical values is less than satisfactory.

Fu and Li (1993) used simulated data with known values of θ and n to locate appropriate quantiles as estimates for the critical value of the statistic. Then for each value of n , they took the most extreme of these critical values over all θ 's in the interval $[2, 20]$. The effect of this technique is to reject only when the data cannot be explained by any value of θ in this interval. The interval $[2, 20]$ for θ was chosen somewhat arbitrarily to represent "most cases of interest."

While Fu and Li's approach is an improvement over Tajima's, there are still some problems remaining. First, the critical values are not applicable when the true value of θ is not in $[2, 20]$, and we cannot know, for a given set of data, whether this is the case. Since θ is a per-locus value, it changes with the number of nucleotides being sequenced, as well as with the underlying mutation rate. Thus it is difficult to justify why θ would have to be confined to this range. The test may falsely reject when θ is not in this interval. Further, their technique does not take into account the information about θ inherent in the data. We will attempt to address these problems below. Since each test statistic has a discrete distribution, any non-randomized test will not precisely achieve the desired level.

The problem of the unknown parameter θ may be addressed using the technique proposed by Berger and Boos (1994). Ideally, we would like to reject only if the data cannot be explained by any positive value of θ . For a test of level α this would mean choosing critical values $[D_L, D_U]$ for D such that

$$\sup_{\theta \in [0, \infty)} [Pr_{\theta}(D \leq D_L) + Pr_{\theta}(D \geq D_U)] \leq \alpha \quad (17)$$

However, we cannot perform simulations for infinitely many values of θ , nor is it reasonable to do so since extremely large values of θ are unlikely. Instead, for some small number $\beta < \alpha$, we use the data to estimate C_{β} , a $1 - \beta$ confidence interval for θ , and require critical values to satisfy

$$\sup_{\theta \in C_{\beta}} [Pr_{\theta}(D \leq D_L) + Pr_{\theta}(D \geq D_U)] \leq \alpha - \beta. \quad (18)$$

For each θ in a grid covering C_{β} we estimate level $(\alpha - \beta)$ critical values $[D_L^{\theta}, D_U^{\theta}]$ using the quantiles of neutral data for that θ . We take the most extreme of those critical values over all θ in C_{β} :

$$D_L = \inf_{\theta \in C_{\beta}} D_L^{\theta}, \quad D_U = \sup_{\theta \in C_{\beta}} D_U^{\theta}. \quad (19)$$

The result is a level α test, as shown by Berger and Boos (1994). This approach is similar to that of Fu and Li, except that instead of arbitrarily using the interval $[2, 20]$ we use an interval that reflects our knowledge of θ for this data set. This has the advantage of giving us a test with known level for any value of the unknown parameter θ .

To construct this $1 - \beta$ confidence interval for θ , we use the exact distribution for S given θ , as given by Tavaré (1984). We wish to find a two-sided interval $C_{\beta} = [\theta_L, \theta_U]$ such that, for a particular observation $S = s$, and for fixed n ,

$$P(S \geq s \mid \theta = \theta_L) = \beta/2 \quad (20)$$

$$P(S \leq s \mid \theta = \theta_U) = \beta/2. \quad (21)$$

The cumulative distribution function for S given θ is

$$F(s, n, \theta) = P(S \leq s \mid \theta) = 1 - \sum_{r=1}^{n-1} (-1)^{r-1} \binom{n-1}{r} \left(\frac{\theta}{r+\theta} \right)^{s+1}. \quad (22)$$

So, (20) and (21) may be written as

$$F(s-1, n, \theta_L) = 1 - \beta/2 \quad (23)$$

$$F(s, n, \theta_U) = \beta/2 \quad (24)$$

Thus we must solve (23) and (24) for θ_L and θ_U for the particular values of $S = s$ and n observed in the data. This is computationally intensive for large values of n , and requires high precision to compute accurately in many cases. We used the variable-precision capabilities of the symbolic computation package *Maple* (CHAR *et al.* 1991) to perform the calculations. The results of these computations are given in the Results section. Note that when $S = 0$ is observed it is appropriate to set $\theta_L = 0$ and solve $F(0, n, \theta_U) = \beta$ for θ_U in place of (24).

In summary, there are three distinct steps to computing critical values for the test statistics in this fashion.

1. For the values of n and S required, compute C_β , a $1 - \beta$ confidence region for θ given S .
2. For a grid of θ -values in C_β and for each n , simulate a large number of samples and estimate level- $(\alpha - \beta)$ critical values for each test statistic from the simulated empirical distributions.
3. Take the maximum upper critical value and minimum lower critical value over all values of θ in C_β , for each value of n and S and for each test statistic. This gives critical values of α -level tests for each n and S .

Simulations

To evaluate the power of the statistical tests described above, we require data simulated under a number of different alternative models. The alternatives considered here: a selective sweep event, a population bottleneck, and a subdivided population, represent a few simple deviations from strict neutrality, and are meant as examples rather than as a comprehensive study. Since balancing selection is similar to population subdivision from a coalescent perspective (HUDSON 1990), we expect the results for a subdivided population to be applicable to the corresponding balancing selection alternative as well.

Neutral Simulations: A sample of DNA sequences is generated by simulating a random gene genealogy. There are three components to this genealogy: topology, branch lengths, and mutations. First, a random tree topology is generated for the genealogy. From n individuals in the sample, two are chosen at random to be the first to coalesce. A new individual is designated as their parent, and the process is repeated on the remaining $n - 1$ individuals. The process stops when only one individual, the most recent common ancestor (MRCA) of the entire sample, remains. This gives the topology of a binary tree with n tips. Next, the branch lengths are chosen: t_j , the time (in units of $2N$ generations) during which there are exactly j lineages, is an exponentially distributed random variable with mean $1/\binom{j}{2}$ as described in the Introduction. These two steps define a tree such as that shown in Figure 1. Finally, mutations are added to the tree. The number of mutations S that have occurred during the history of the sample is generated as a Poisson-distributed random variable with mean $\theta T_{tot}/2$. For each mutation, the branch of the tree on which it occurred is chosen randomly, where the relative probability of each branch is proportional to its length. The mutation is transmitted to each offspring descended from that branch. Thus each individual is assigned a “sequence” of nucleotides designated, for example, $-+--+--$, where “-” indicates that the nucleotide is identical to the ancestral sequence at that site, and “+” indicates a mutation. Under the infinite-sites model, each mutation is assumed to take place at a distinct nucleotide site, and thus each sequence generated is composed only of polymorphic or segregating sites.

Selective Sweep Simulations: A highly favorable mutation with selective advantage s and dominance h that occurs at a time t_s is assumed to sweep through the population and reach fixation in a deterministic fashion, such that the proportion $x(t)$ of individuals carrying the mutation at time t follows

$$\dot{x}(t) = \frac{2Nsx(1-x)[x+h(1-2x)]}{1+s[x^2+2hx(1-x)]}, \quad x(t_s) = \frac{1}{2N} \quad (25)$$

This result can be found in (MAYNARD SMITH and HAIGH 1974, equation (18)). (We have inserted a multiplicative factor of $2N$ to correct for the measurement of time in units of $2N$ generations.)

The selective sweep alters the coalescent process by reducing the effective population size of the parental generation at time t from $2N$ to $2Nx(t)$, since only genes carrying the selected mutation may be chosen as ancestors of the sample. The per-generation coalescent probabilities change from $\frac{\binom{j}{2}}{2N}$ to $\frac{\binom{j}{2}}{2Nx(t)}$. Thus the total size of the tree is reduced, and the effect of the selective sweep is to reduce variation at and around the selected locus.

To generate coalescent times under a sweep we generate times according to the neutral model, and then scale them appropriately, as described below. This approach was suggested to us by R. R. Hudson; also see (GRIFFITHS and TAVARÉ 1994, equation (3)). To convert a time from one time scale to another we must perform a change of variables. Suppose U is a time measured in units of $2Nx(t)$ at time t . We wish to convert this time U back into the standard units of $2N$ generations. The instantaneous change of variables at time t is

$2Nx(t)du = 2Ndt$, where dt is the interval in regular units $2N$ and du is the time interval in units $2Nx(t)$. This becomes

$$\frac{dt}{x(t)} = du, \quad (26)$$

and thus, if T represents the same time as U but in regular units, we integrate over the whole interval to obtain

$$\int_0^T \frac{dt}{x(t)} = \int_0^U du = U \quad (27)$$

Therefore, to generate a coalescent time T under a selective sweep described by $x(t)$, we generate a time U under the neutral model, and then find T which solves (27). This is done for each coalescent time in a tree, to generate a coalescent tree for a selective sweep. Note that the MRCA of the sample has to occur more recently than t_s in this model. If the sweep began so recently that the selected allele has not yet completely reached fixation, we assume that the alleles in the sample are all descendants of the selected allele.

Our model of the selective sweep is defined in terms of four parameters: h , s , N , and its starting time t_s . For this study, we chose to fix $h = 0.5$, $N = 10^6$, and $s = 10^{-4}$, and allow t_s to vary over the range zero to four (in units of $2N$, back in time from the present). This is relatively weak selection on a co-dominant allele; for comparison we also performed some simulations with $s = 10^{-2}$. For various combinations of t_s , n , and θ , 1000 samples were generated and the proportion of rejections for each of the tests recorded.

A selective sweep is expected to reduce polymorphism at linked sites, since any observed polymorphism must be the result of mutations that have occurred since the sweep. These newly arisen mutations will at first be rare, and will increase in frequency as the time since the sweep increases. Since S takes into account only the number of mutations, while k is also affected by their frequency, it is expected that S will recover more rapidly than k from the effects of a sweep. This will have the effect of reducing the expected value of Tajima's statistic below its neutral expectation of zero. The magnitude of this reduction has not been predicted by theory, and is one of the subjects of the present investigation.

Population Bottleneck Simulations: A population bottleneck is assumed to occur when the population, originally of size $2N$, is suddenly reduced to a fraction f of its former size for a length of time l , then instantaneously regains its initial size. Let t_b be the amount of time in units of $2N$ generations since the bottleneck ended, so that it began at a time $t_b + l$ which is further from the present time zero. Coalescent times under this model are obtained by scaling neutral coalescent times as with the selective sweep. In this case, the changing population size $2Nx(t)$ is a step function rather than a smooth curve as it was for the sweep, so the integration in (27) is easy. We generate a time u_j under the neutral model, and then use as the coalescent time t_j given by

$$t_j = \begin{cases} u_j, & u_j < t_b \\ t_b + (u_j - t_b)f, & t_b < u_j < t_b + \frac{l}{f} \\ u_j - (\frac{1}{f} - 1)l, & t_b + \frac{l}{f} < u_j. \end{cases} \quad (28)$$

This is equivalent to the following probability density for the coalescent times t_j :

$$t_j \sim \begin{cases} pe^{-pt_j}, & t_j < t_b \\ \frac{p}{f} e^{-\frac{p}{f}t_j} e^{pt_b(\frac{1}{f}-1)}, & t_b < t_j < t_b + l \\ pe^{-pt_j} e^{-pl(\frac{1}{f}-1)}, & t_b + l < t_j \end{cases} \quad (29)$$

where $p = \binom{j}{2}$. The density can be derived by considering the per-generation probabilities of coalescence during the three stages of the bottleneck.

For purposes of this study, we kept f fixed at 0.01, l fixed at 0.1, and varied t_b , the time since the bottleneck ended, from zero to five. These are bottlenecks of the same severity but lasting ten times the length of those considered in (TAJIMA 1993). The fraction rejected out of 1000 simulations was recorded.

A population bottleneck is expected to reduce polymorphism throughout the genome, since a drastic reduction in population size is likely to eliminate many rare variants. As in the case of the selective sweep, most of the polymorphism will be a result of new mutations, which will be rare. Thus a reduction of unknown magnitude in the expectation of Tajima's statistic is predicted (TAJIMA 1989b).

Subdivided Population Simulations: The third alternative modeled was a subdivided population with no migration. We expect the results of this model to apply to balanced polymorphism as well, since the two are similar from a coalescent perspective. We start with an ancestral population size of $2N_{AB}$. At a certain time t_m this population is assumed to split into two isolated populations A and B , of size N_A and N_B respectively, which evolve independently from then on. Here, the sample of n alleles consists of n_A alleles of type A , and n_B of type B , with $n = n_A + n_B$.

The coalescent tree for such a subdivided population is generated in the following manner. As usual we work backwards in time from the present to the time of the MRCA of the sample. Let j_A and j_B be the number of lineages of type A and B remaining at any given time. Initially, we let $j_A = n_A$ and $j_B = n_B$, and at each coalescent event, one of them is decremented. We need to know the distribution of the time back to the next coalescent event.

In the subdivided model, the coalescent probabilities before and after the population split are different. When the two populations are disjoint, the probability per generation that two A individuals coalesce is $p_A = \binom{j_A}{2} / 2N_A$ while for the B population it is $p_B = \binom{j_B}{2} / 2N_B$. The probability that both populations will coalesce in the same generation is negligible ($O(\frac{1}{N^2})$) compared to p_A and p_B , and so the per-generation probability of a coalescent event in either population is approximately $p_1 = p_A + p_B$ while the two populations are disjoint. When the two populations are mixed, we have a single population of size $2N_{AB}$, with $j_A + j_B$ sample lineages present. Thus the per-generation probability of coalescence for the mixed population is $p_2 = \binom{j_A + j_B}{2} / 2N_{AB}$.

As before, t_i is the time during which there are exactly i lineages of any type present. Let $S_j = \sum_{i=j}^n t_i$, with $S_{n+1} = 0$. S_j keeps track of the total time generated so far. To generate

the time $t_{j_A+j_B}$ to the next event, it is necessary to know the relationship between $S_{j_A+j_B+1}$ and t_m . In particular, if $S_{j_A+j_B+1} > t_m$ then we have passed the subdivision point and we may generate subsequent times $t_{j_A+j_B}$ simply as exponentially distributed random variables with parameter p_2 . On the other hand, suppose $S_{j_A+j_B+1} < t_m$, say $t_m - S_{j_A+j_B+1} = M > 0$. Then the time $t_{j_A+j_B}$ generated could be less than or greater than M . The probability of coalescence after a given time $t < M$ is $(1 - p_1)^t p_1 \approx p_1 e^{-p_1 t}$. But for a time $t > M$, the probability of coalescence is $(1 - p_1)^M (1 - p_2)^{t-M} p_2 \approx p_2 e^{-p_2 t} e^{M(p_2 - p_1)}$. Thus

$$t_{j_A+j_B} \sim \begin{cases} p_1 e^{-p_1 t} I_{[t < M]} + p_2 e^{-p_2 t} e^{M(p_2 - p_1)} I_{[t > M]}, & M > 0 \\ p_2 e^{-p_2 t}, & M < 0 \end{cases} \quad (30)$$

where I is an indicator function and $M = t_m - S_{j_A+j_B+1}$.

Once a time has been generated from this mixture of exponentials, two individuals must be chosen to coalesce at that time. If the total time $S_{j_A+j_B+1}$ is still less than t_m , then we must choose between group A and group B with relative probabilities p_A and p_B . If the time is greater than t_m , we have only one group. Once the group is chosen, two of the appropriate group are selected at random, and the corresponding j is decremented. The process is repeated until only one individual remains.

When a population is subdivided, the average pairwise difference k is inflated relative to the total number of mutations S , because of the large divergence between subpopulations. Thus the qualitative expectation is that D will have a positive mean in this situation. As with the selective sweep and bottleneck, we chose time since the subdivision event as the primary variable to investigate, fixing $N_A = N_B = N_{AB}/2$, $n_A = n_B = 25$, and $\theta = 20$.

RESULTS

Results of Neutral Simulations

Simulations of the null hypothesis were used to provide new critical values for the test statistics. Our technique uses confidence intervals for θ given S , as described in the Methods section. These $1 - \beta$ confidence intervals were computed using 40 digits of accuracy to solve equations (23) and (24). Tabulation here of these confidence intervals for different values of n , S , and β would be prohibitive, so we show only a sample: the case $n = 50$, $\beta = 0.01$ in Table 1. This table shows, for example, that if $S = 23$ is observed from a sample of size $n = 50$, and the neutral model holds, then with 99% certainty θ is between 2 and 12.5. For other values of n and β , θ_L and θ_U may be closely approximated by linear functions of S , especially when S is large. For example, when $n = 20$, $C_{0.01}$ is approximately $[0.121S - 0.481, 0.709S + 2.858]$ and $C_{0.001}$ is approximately $[0.094S - 0.473, 0.904S + 4.418]$. The coefficients of these linear approximations are given in Table 2, and can be used to approximate the values corresponding to Table 1 for other values of n and β .

Table 3 shows tables of level 0.05 critical values for Tajima's test for a range of S values, for $n = 10, 20, 50, 100$, using $\alpha = 0.05$ and $\beta = 0.01$. For comparison, the values from the beta distribution (TAJIMA 1989a) are also shown. Corresponding values for D^* and F^* are given in Tables 4 and 5, along with the values that assume $\theta \in [2, 20]$ from (FU and LI 1993).

There is no simple pattern to the way in which the new critical values differ from those of the beta distribution. Generally speaking, for small n the beta distribution values are too large, while for larger n the beta distribution values are too small. The important difference is that the new values are based on a sound statistical framework that does not depend on fitting the statistic to a particular distribution, as Tajima did, or on the true value of θ being between two and 20, as Fu and Li assumed.

The size of these tests (the probability of rejecting when the neutral model is true), based on the new critical values, was estimated by applying each test to 10,000 simulated neutral data sets for each value of θ . The number of false rejections was computed (data not shown). The size for most values of θ is between 3% and 4%, out of a maximum of $\alpha = 5\%$. This shortfall is attributable to three factors. First, since the statistics have discrete distributions we cannot expect to precisely achieve the desired level with any non-randomized test. Second, there is some error in estimating the $(\alpha - \beta)$ critical values using the empirical quantiles, since we used a finite number of simulations (10,000). This source of error could be diminished, though not eliminated, by using a larger number of simulations. Third, the Berger and Boos confidence interval procedure is conservative; using it may reduce the size of the test by as much as β . Thus, the critical values might be improved by using a smaller value of β .

Results of Selective Sweep Simulations

The effect of a selective sweep on Tajima's D statistic is shown in Figure 2 for two different strengths of selection: (a) $s = 10^{-4}$ (weak); (b) $s = 10^{-2}$ (stronger). The horizontal axis is t_s , the time since the sweep began. The solid curve is the median of Tajima's D over 1000 simulations with $n = 50$ and $\theta = 20$, while the dashed lines are the 2.5 and 97.5 percentiles. The horizontal lines are the critical values from Tajima's (1989a) beta distribution. The expected trend towards more negative values of D is observed, but except in a particular time window, the reduction is not large enough to make rejection very probable. There is also a pronounced decrease in the variance of the distribution even when the sweep is very ancient. When t_s is very large (six to eight), the percentile curves eventually level off close to the critical values. When t_s is very small, the selective sweep is so recent that there are few if any segregating sites in the sample, with the result that D is close to zero with high probability. (Note that D cannot be computed when $S = 0$.) Comparing Figures 2(a) and 2(b) shows the effect of the strength of selection: stronger selection results in a more immediate decrease in the expected value of D . Note, however, that the length of time (approximately $2\ln(2N)/Ns$ when $h = 0.5$) it takes the sweep to complete must be taken into account; this is approximately 0.3 and 0.003 for the weak and strong cases, respectively (shown inset in Figure 2). Thus in (a), when $t_s < 0.3$ the sweep is still in progress at time 0, and since all sampled individuals must be descended from a single individual at time t_s , there is almost no variation. In (b), on the other hand, the sweep is virtually instantaneous compared to the scale shown (though it takes 6000 generations), so the sample has had more time to recover variation after the sweep.

The power of Tajima's D test against the selective sweep alternative is shown in Figure 3:

(a) $\theta = 10$, $s = 10^{-4}$; (b) $\theta = 20$, $s = 10^{-4}$; (c) $\theta = 50$, $s = 10^{-4}$; (d) $\theta = 20$, $s = 10^{-2}$. The horizontal axis is t_s as in Figure 2, but note that the scale is enlarged. The different curves are for different values of n as labeled on the graphs. Figure 3 shows that the sample size has a profound effect on the power to reject. While a sample of size 50 or 100 can give a substantial power, no significant result can be expected from a sample size of 10 in most cases. It appears that even with large sample sizes, it is only possible to detect selective sweeps that occurred in a specific window of time. For example, if $n = 100$ and $\theta = 20$, Tajima's test will reject with probability 90% only if the sweep (weak selection) began between $t_s = 0.2$ and $t_s = 0.3$, which, with $N = 10^6$ corresponds to between 400,000 and 600,000 generations ago. It must be emphasized that these results apply only to the particular model of sweep and the parameter values ($s = 10^{-4}$, $h = 0.5$) used in the simulation. For clarity, the graphs in Figure 3 are shown with t_s in the range zero to one. However, simulations were actually performed for t_s as large as four. It can be seen in the figures that the power drops well below the neutral expectation of 0.05 when t_s is close to one. In fact, for t_s from one to four the data do not behave neutrally. For these t_s , the sweep was long enough ago that new mutations have had a chance to reach intermediate frequency in the population, but polymorphism is still quite reduced. In other words, the difference between the expectations of k and S/a_n is fairly small, but the variance of that difference is still reduced well below one. This has the paradoxical result of making the test less likely to reject under the alternative than under the null hypothesis, when t_s takes on intermediate values. In other words, these tests are biased.

Figure 4 shows the power of all nine tests against the selective sweep alternative when $n = 50$ and $\theta = 20$. Among all the tests considered, Tajima's test showed the most power to reject a selective sweep for each value of n and θ we simulated. The tests T_2 and F_3 were almost as powerful as D , and were more powerful than Fu and Li's F^* and D^* tests. Although Tajima's test statistic does lack power in many cases, it appears to be the most powerful test of this class against the selective sweep alternative as modeled here.

Results of Population Bottleneck Simulations

The results for the population bottleneck are summarized in Figures 5 and 6. Each figure represents a bottleneck lasting 0.1 (units $2N$ generations) and dropping to 1% of its original size. The horizontal axis in each case is the time t_b since the bottleneck, and each data point is based on 1000 simulations. Figure 5 shows the median and 2.5 and 97.5 percentiles of Tajima's statistic D , versus t_b , for $n = 50$ and $\theta = 20$. Figure 6 shows the fraction rejected by Tajima's test for the cases (a): $\theta = 10$; (b): $\theta = 20$; (c): $\theta = 50$, and by Fu and Li's F^* test for the case (d): $\theta = 20$. The results are similar to those for the selective sweep. A bottleneck is only likely to be detected if it is very recent, and if the sample size is large. Again, Tajima's test performs the best of all the tests considered. The similarity of the results to those for a selective sweep is to be expected, since the effect on the coalescent process of the two situations is similar.

Results of Population Subdivision Simulations

Population subdivision has an effect opposite to that of a selective sweep on the statistics being studied. A subdivided population results in a higher value of k than would be expected under neutrality, while the effect on S is less severe. Thus population subdivision tends to produce positive values of the test statistics D , F^* , and D^* . The more ancient the division, the greater this effect becomes. A plot of the median and 0.025 and 0.975 percentiles of Tajima's D against the time of separation t_m , for a sample size of 50 ($n_A = n_B = 25$) and $\theta = 10$, with $N_{AB} = N_A + N_B$, is shown in Figure 7. Power curves for all nine tests are shown in Figure 8. It can be seen from this figure that the probability of detecting this type of population subdivision with these tests is quite small unless the division is fairly ancient. Again, Tajima's D is the most powerful test against this alternative, with T_2 having almost identical power to D , and Fu and Li's F^* the next most powerful. The above results were given for $n_A = n_B = 25$. When we choose $n_A \neq n_B$, (e.g. $n_A = 10, n_B = 40$) the power is even less, with all other parameters held fixed (results not shown).

Some Comments on Sample Size

We have shown that sampling a greater number of individuals increases the power of the test. But, sampling longer sequences (effectively increasing θ) should also increase the power. Which is better: longer sequences or more individuals? To answer this question, we must assign a relative cost to these two options. Let us assume that the cost per nucleotide sequenced is the same whether that nucleotide comes from a new individual, or from extending the sequenced region. This ignores costs associated with both the acquisition and preparation of a new individual, and the cloning of longer regions. Further suppose that the per-locus mutation rate is proportional to the length of the sequence, so that doubling the number of bases doubles θ . Under these assumptions, the cost is proportional to the product of n and θ . Therefore, we compare power curves where the product of n and θ is the same.

In Figure 9, we show the power of Tajima's test against a selective sweep for the product $n\theta = 200, 500$, and 1000. Note that in this context, increasing θ means increasing the size of the region examined (and thus μ) for a given N . These results show that, against the alternative of a selective sweep, it is better to sequence more individuals than more sites, so long as the number of sites is not too small. For example, against the selective sweep alternative as modeled here, Tajima's test is always more powerful when $n = 20$ and $\theta = 10$ than when $n = 10$ and $\theta = 20$. Similar results hold for other tests.

DISCUSSION

The new method of calculating critical values for the class of tests presented here allows us to eliminate from the null hypothesis the requirement (Fu and Li 1993) that θ is between two and 20, at the expense of extra computation. If the true value of θ for a studied locus is indeed in that range, there is very little difference between the two methods. However, our method has the advantage that rejection cannot be explained by a too-small or too-large θ .

If Fu and Li's published critical values are used, it should be with the understanding that the true level of the test should have added to it the probability that θ is not in that range. For Tajima's test, our critical values are a clear improvement over the beta distribution method. In many cases, Tajima's published values are too conservative, with the result that rejection is almost impossible. Our new critical values result in a more powerful test.

As an alternative method of examining the behavior of D when θ is unknown, other authors (HUDSON 1993; BRAVERMAN *et al.* 1995) have suggested sampling from the conditional distribution of D given S , where S is obtained from the data set to be tested. With S fixed, D is simply a linear transformation of k , and may therefore have a smaller variance under neutrality since the contribution of S to the variance is eliminated. Both methods choose a genealogy at random, but their method fixes S for all genealogies, whereas we fix θ and from this generate a value of S based on the total time in the genealogy. The contrast between the two methods of generating S is most evident when simulating data from alternative hypotheses in order to estimate power. The two methods represent two different views of the power of a test: as a function of the parameter θ , and as a function of the statistic S . We investigate the behavior of D after a selective sweep, with several different, but fixed, mutation rates. Their method examines the effect of a selective sweep after which a fixed number of mutations has occurred.

Among all the tests considered, Tajima's test (with the new critical values) was the most powerful against the specific alternatives we simulated. Certainly we cannot extrapolate from this to say it is more powerful against all possible alternatives and parameter values. However, since the chance of spurious rejection increases with the number of tests performed, we want to perform as few tests as possible. Therefore we want to perform only the test with the greatest chance of rejection. In the absence of other evidence, that would appear to be Tajima's test. Using a different model of recurrent hitchhiking under very recent, strong selection, Braverman *et al.* (1995) also found Tajima's D (conditional on S) to be more powerful than Fu and Li's D^* . The new test statistics described above do not perform as well as Tajima's test, although they do have more power than Fu and Li's tests in many cases. Thus we do not recommend their use.

Our results indicate that sample sizes of at least 50 are typically necessary to achieve any reasonable power. Most sample sizes for sequence data seen in the literature are much smaller than this. However, even for large sample sizes, the probability of detecting a selective sweep that is not recent is quite small.

We have shown that a negative expected value of Tajima's D is in fact observed at linked neutral sites after the selective fixation of an advantageous mutation in a model with no recombination. It is also apparent that the ability to detect the selective sweep by either Tajima's (1989a) or Fu and Li's (1993) test statistics is strongly influenced by the strength of selection and by the amount of time since the selective sweep occurred. With an effective population size of 10^6 , selective sweeps of co-dominant mutations with a selective advantage of 10^{-4} result in distributions of variation that are unlikely to be found incompatible with a neutral model using these tests. Increasing the selective advantage 100-fold to 10^{-2} leads to a certain increase in the power of available tests. Nonetheless, there exists a defined window over which the tests have reasonable statistical power to reject the neutral model. For strong

selection, this window appears to be from roughly 100,000 to 400,000 generations. More recent sweeps are undetectable since there has been too little time for sufficient new variants to arise, and if a sweep is too distant the power to reject neutrality drops precipitously as new neutral variants accumulate.

These results suggest that while recent genetic hitchhiking driven by strong selection of the kind modeled here is a somewhat unlikely explanation for reduced levels of variation where a significant Tajima's D test is not observed, e.g., (AGUADÉ *et al.* 1994; BEGUN and AQUADRO 1995), it cannot be ruled out based solely on Tajima's test. Furthermore, less recent sweeps and weaker selection are consistent with the values of Tajima's D that have been observed. The extent to which weaker or more distant selection could result in the observed patterns of data needs further examination. Situations where negative Tajima's D have been observed together with reduced variation do appear consistent with a simple selective sweep model (MARTÍN-CAMPOS *et al.* 1992). The generally low level of power of Tajima's and Fu and Li's test statistics does indicate that other means to distinguish between selective sweeps and background selection should be sought before firm conclusions are drawn. In order to do this, any test will have to take into account more information from the data than just differences between the three summary statistics k , S , and η_s . The apparent contrast between predictions for X-linked versus autosomal gene variation is but one possibility (AQUADRO *et al.* 1994).

This approach to estimating the power of statistical tests should prove useful in investigating many other types of alternatives and statistical tests. For example, it would be useful to know whether the existing tests are able to detect selection against background selection. Tests that use more information from the data, such as outgroups, may be more powerful than the tests studied here. We (with M. J. Ford) are currently undertaking an similar analysis of the properties of the HKA test (HUDSON *et al.* 1987).

ACKNOWLEDGMENTS

We thank R. L. Berger, R. R. Hudson and members of the Aquadro Lab: D. J. Begun, M. J. Ford, M. T. Hamblin, M. W. Nachman and K. S. Phillips, for helpful suggestions. Preliminary discussions on this project involved the above lab members as well as D. Grove, C. McCulloch, and S. Schwager. Some algorithms used in computer programs were based on code originally written by R. R. Hudson. This work was supported in part by NIH grant GM36431 to C.F.A., DOE grant DEFG0293ER61567 to G.A.C., and NSF grant BIR-9113307 to C.F.A. and G.A.C.

APPENDIX

The following are the coefficients of Tajima's and Fu and Li's tests.

$$v_T = \left(\frac{2(n^2 + n + 3)}{9n(n-1)} - \frac{n+2}{a_n n} + \frac{b_n}{a_n^2} \right) / (a_n^2 + b_n) \quad (31)$$

$$u_T = \left(\left(\frac{n+1}{3(n-1)} - \frac{1}{a_n} \right) / a_n \right) - v_T \quad (32)$$

$$v_{D^*} = \left(\frac{b_n}{a_n^2} - \frac{2}{n} \left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n} \right) - \frac{1}{n^2} \right) / (a_n^2 + b_n) \quad (33)$$

$$u_{D^*} = \left(\left(\frac{(n-1)}{n} - \frac{1}{a_n} \right) / a_n \right) - v_{D^*} \quad (34)$$

$$v_{F^*} = \left(\frac{2n^3 + 110n^2 - 255n + 153}{9n^2(n-1)} + \frac{2(n-1)a_n}{n^2} - \frac{8b_n}{n} \right) / (a_n^2 + b_n) \quad (35)$$

$$u_{F^*} = \left(\left(\frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)} \right) / a_n \right) - v_{F^*} \quad (36)$$

The following are the coefficients of the new statistical tests described in the Methods section.

$$v_{T_2} = \frac{\left[2(n^2 + n + 3) - \frac{9(n-1)(n+2)}{a_n} + \frac{9n(n-1)b_n}{a_n^2} \right]}{11n^2 - 7n + 6} \quad (37)$$

$$u_{T_2} = \frac{n+1}{3(n-1)} - 1/a_n - \frac{n+1}{3(n-1)} v_{T_2} \quad (38)$$

$$v_{T_3} = \frac{\left[\frac{2(n^2+n+3)}{9n} - \frac{(n+2)(n-1)}{na_n} + \frac{b_n(n-1)}{a_n^2} \right]}{(2a_n + n + 1)} \quad (39)$$

$$u_{T_3} = \frac{n+1}{3n} - \frac{n-1}{na_n} - v_{T_3} \quad (40)$$

$$v_{D_2^*} = \frac{9(n-1)}{11n^2 - 7n + 6} \left[\frac{nb_n}{a_n^2} - 2 \left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n} \right) - \frac{1}{n} \right] \quad (41)$$

$$u_{D_2^*} = \frac{n-1}{n} - \frac{1}{a_n} - \frac{(n+1)}{3(n-1)} v_{D_2^*} \quad (42)$$

$$v_{D_3^*} = \frac{n-1}{2a_n + n + 1} \left[\frac{b_n}{a_n^2} - \frac{2}{n} \left(1 + \frac{1}{a_n} - a_n + \frac{a_n}{n} \right) - \frac{1}{n^2} \right] \quad (43)$$

$$u_{D_3^*} = \frac{(n-1)^2}{n^2} - \frac{n-1}{na_n} - v_{D_3^*} \quad (44)$$

$$v_{F_2^*} = \frac{\left[2n^3 + 110n^2 - 255n + 153 + 18(n-1)^2 a_n - 72n(n-1)b_n \right]}{n(11n^2 - 7n + 6)} \quad (45)$$

$$u_{F_2^*} = \frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n(n-1)} - \frac{n+1}{3(n-1)}v_{F_2^*} \quad (46)$$

$$v_{F_3^*} = \frac{\left[\frac{2n^3+110n^2-255n+153}{9n^2} + \frac{2(n-1)^2a_n}{n^2} - \frac{8b_n(n-1)}{n} \right]}{2a_n + n + 1} \quad (47)$$

$$u_{F_3^*} = \frac{4n^2 + 19n + 3 - 12(n+1)a_{n+1}}{3n^2} - v_{F_3^*} \quad (48)$$

LITERATURE CITED

- AGUADÉ, M., W. MEYERS, A. D. LONG, and C. H. LANGLEY, 1994. Reduced DNA sequence polymorphism in the *su(s)* and *su(w^a)* regions of *Drosophila melanogaster* as revealed by SSCP and stratified DNA sequencing. *Proc. Natl. Acad. Sci. USA* **91**: 4658–4662.
- AGUADÉ, M., N. MIYASHITA, and C. H. LANGLEY, 1989. Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. *Genetics* **122**: 607–615.
- AQUADRO, C. F., D. J. BEGUN, and E. C. KINDAHL, 1994. Selection, recombination, and DNA polymorphism in *Drosophila*. In GOLDING, B., editor, *Non-Neutral Evolution: Theories and Molecular Data*, chapter 4, pages 46–56. Chapman and Hall, New York.
- BEGUN, D. J. and C. F. AQUADRO, 1991. Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: Evidence for genetic hitchhiking of the *yellow-achaete* region. *Genetics* **129**: 1147–1158.
- BEGUN, D. J. and C. F. AQUADRO, 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.
- BEGUN, D. J. and C. F. AQUADRO, 1995. Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. *Molecular Biology and Evolution*. in press.
- BERGER, R. L. and D. D. BOOS, 1994. P-values maximized over a confidence set for a nuisance parameter. *Journal of the American Statistical Association* **89**(427): 1012–1016.

- BERRY, A. J., J. W. AJIOKA, and M. KREITMAN, 1991. Lack of polymorphism in the *Drosophila* fourth chromosome resulting from selection. *Genetics* **129**: 1111–1117.
- BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY, and W. STEPHAN, 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*. in press.
- CHAR, B. W., K. O. GEDDES, G. H. GONNET, B. L. LEONG, M. B. MONAGAN, and S. M. WATT, 1991. *Maple V Library Reference Manual*. Springer-Verlag.
- CHARLESWORTH, B., 1994. The effect of background selection against deleterious mutations on weakly-selected, linked variants. *Genetical Research, Cambridge* **63**: 213–227.
- CHARLESWORTH, B., M. T. MORGAN, and D. CHARLESWORTH, 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289–1303.
- FU, Y.-X. and W.-H. LI, 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- GRIFFITHS, R. C. and S. TAVARÉ, 1994. Ancestral inference in population genetics. *Statistical Science* **9**(3): 307–319.
- HUDSON, R. R., 1990. Gene genealogies and the coalescent process. In ANOTOVICS, J. and D. FUTUYAMA, editors, *Oxford Surveys in Evolutionary Biology*, volume 7, pages 1–44. Oxford University Press, Oxford.
- HUDSON, R. R., 1993. The how and why of generating gene genealogies. In TAKAHATA, N. and A. G. CLARK, editors, *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, pages 23–36. Sinauer Associates, Inc., Sunderland, MA.
- HUDSON, R. R., M. KREITMAN, and M. AGUADÉ, 1987. A test of neutral evolution based on nucleotide data. *Genetics* **116**: 153–159.
- KAPLAN, N. L., R. R. HUDSON, and C. H. LANGLEY, 1989. The “hitchhiking effect” revisited. *Genetics* **123**: 887–899.
- MARTÍN-CAMPOS, J. M., J. M. CAMERON, N. MIYASHITA, and M. AGUADÉ, 1992. Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. *Genetics* **130**: 805–816.
- MAYNARD SMITH, J. and J. HAIGH, 1974. The hitch-hiking effect of a favourable gene. *Genetical Research, Cambridge* **23**: 23–35.
- STEPHAN, W. and C. H. LANGLEY, 1989. Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermillion* and *forked* loci. *Genetics* **121**: 89–99.

- TAJIMA, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- TAJIMA, F., 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- TAJIMA, F., 1989b. The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.
- TAJIMA, F., 1993. Measurement of DNA polymorphism. In TAKAHATA, N. and A. G. CLARK, editors, *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, pages 37–59. Sinauer Associates, Inc., Sunderland, MA.
- TAVARÉ, S., 1984. Line-of-descent and genealogical processes and their applications in population genetics models. *Theoretical Population Biology* **26**: 119–164.
- WATTERSON, G. A., 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7**: 256–276.
- WIEHE, T. H. E. and W. STEPHAN, 1993. Analysis of a genetic hitchhiking model and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution* **10**: 842–854.

Figure 1: An example of a coalescent tree for a sample of five genes.

Figure 2: The effect of a selective sweep on Tajima's D statistic: the median and 2.5 and 97.5 percentiles versus the time t_s since the sweep began. Horizontal lines are critical values for rejection. Each data point is based on 1000 simulations of a selective sweep with parameters $\theta = 20$, $n = 50$, $h = 0.5$, and $N = 10^6$: (a) $s = 10^{-4}$; (b) $s = 10^{-2}$. The length of time it takes the selected allele to reach fixation is also depicted (inset); a sweep beginning at t_s ends at the given distance to the left of t_s .

Figure 3: Power of Tajima's D against a selective sweep versus the time since the sweep began for $n = 10, 20, 50$, and 100 : (a) $\theta = 10$; (b) $\theta = 20$; (c) $\theta = 50$; (d) $\theta = 20$. Each data point is based on 1000 simulations of a sweep with parameters $h = 0.5$, $s = 10^{-4}$, and $N = 10^6$, except (d), which uses $s = 10^{-2}$.

Figure 4: Power of all nine statistical tests against a selective sweep versus the time t_s since the sweep began for $n = 50$ and $\theta = 20$. Each data point is based on 1000 simulations of a sweep with parameters $h = 0.5$, and $N = 10^6$: (a) $s = 10^{-4}$; (b) $s = 10^{-2}$.

Figure 5: The effect of a population bottleneck on Tajima's D statistic. Shown are the median (solid line) and 2.5 and 97.5 percentiles (dashed lines) versus the time t_b since the bottleneck ended. Horizontal lines are critical values for rejection. Each point is based on 1000 simulations of a population bottleneck with parameters $\theta = 10$, $n = 50$, $f = 0.01$ and $l = 0.1$.

Figure 6: Power of statistical tests against a population bottleneck versus the time t_b since the bottleneck ended for $n = 10, 20, 50$, and 100 . The tests are (a): D , $\theta = 10$; (b): D , $\theta = 20$; (c): D , $\theta = 50$; (d) F^* , $\theta = 20$. Each data point is based on 1000 simulations of a population bottleneck with parameters $f = 0.01$ and $l = 0.1$.

Figure 7: The effect of a subdivided population on Tajima's D : the median (solid line) and 2.5 and 97.5 percentiles (dashed lines) versus time of separation t_m . Horizontal lines are critical values for rejection. Each point is based on 1000 simulations of population subdivision with parameters $\theta = 10$, $n = 50$, $n_A = n_B = 25$.

Figure 8: Power of all nine tests against population subdivision. Fraction rejected versus time of separation t_m . Based on 1000 simulations of population subdivision with $n = 50$, $n_A = n_B = 25$, and $\theta = 10$.

Figure 9: Power of Tajima's test against a selective sweep versus time since the sweep. Each plot is for a constant value of the product of n and θ : (a) $n\theta = 200$; (b) $n\theta = 500$; (c) $n\theta = 1000$. Each data point is based on 1000 simulations of a selective sweep with parameters $h = 0.5$, $s = 10^{-4}$, and $N = 10^6$.

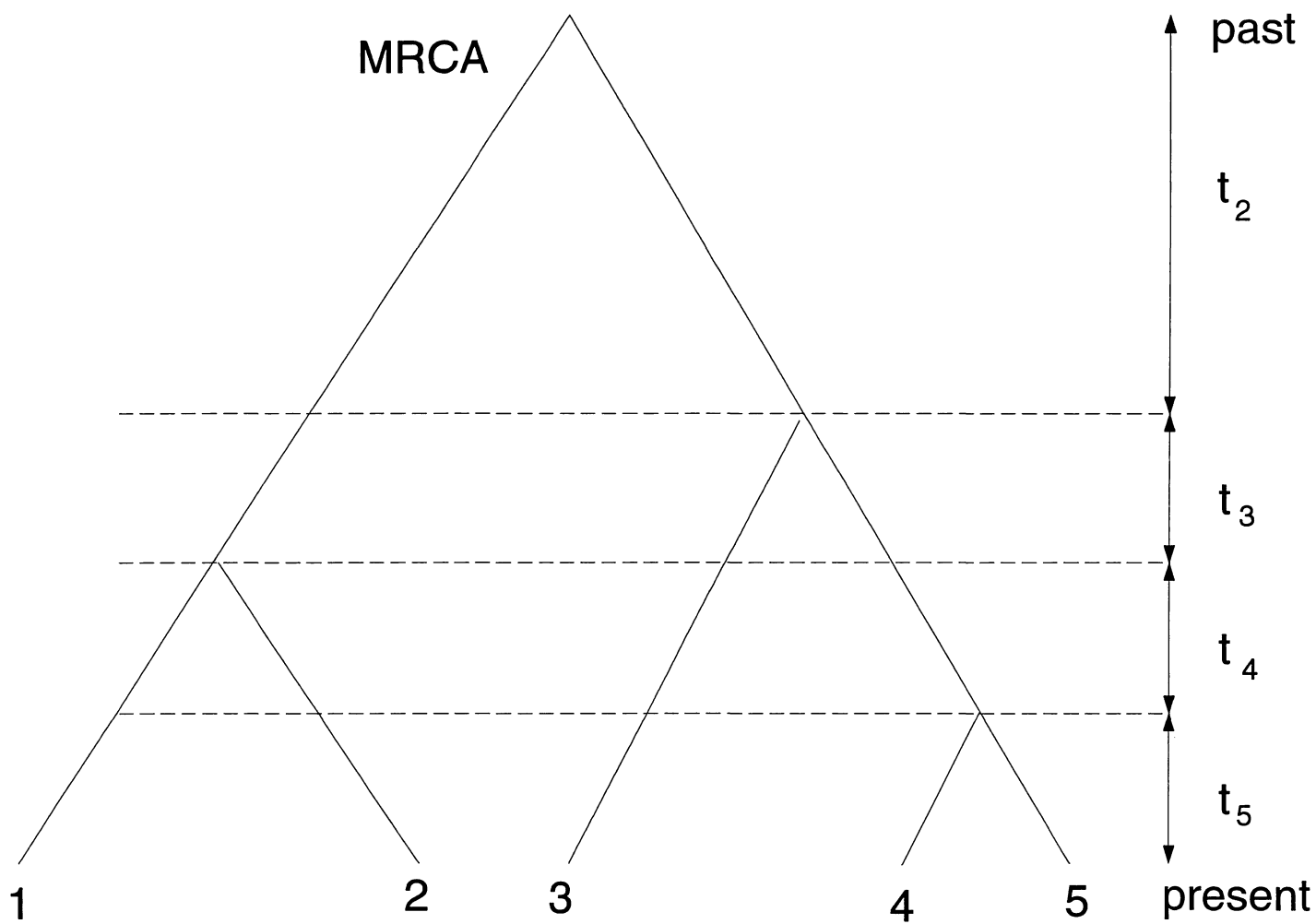


Figure 1

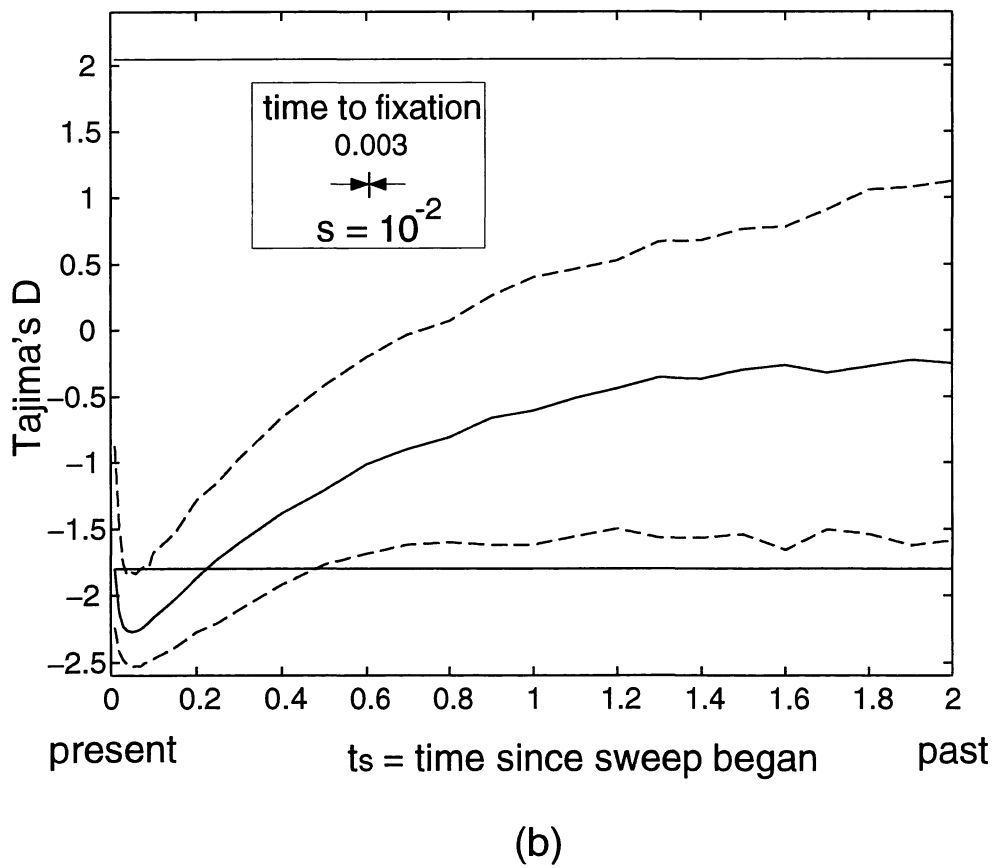
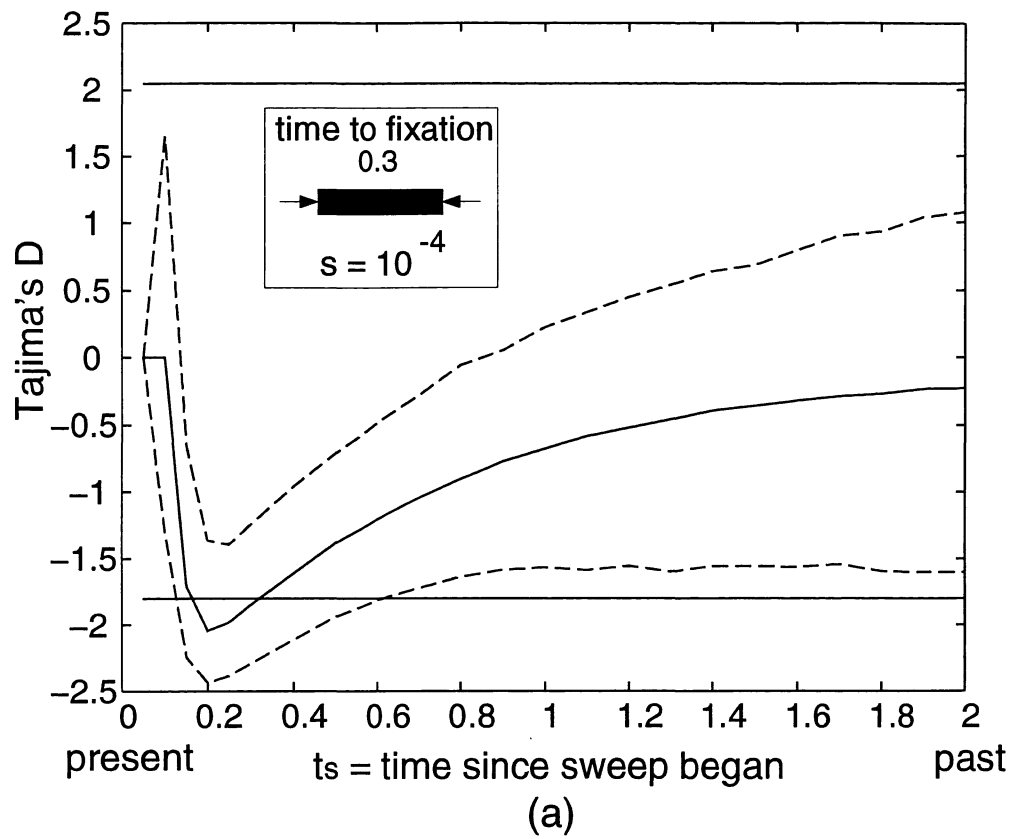
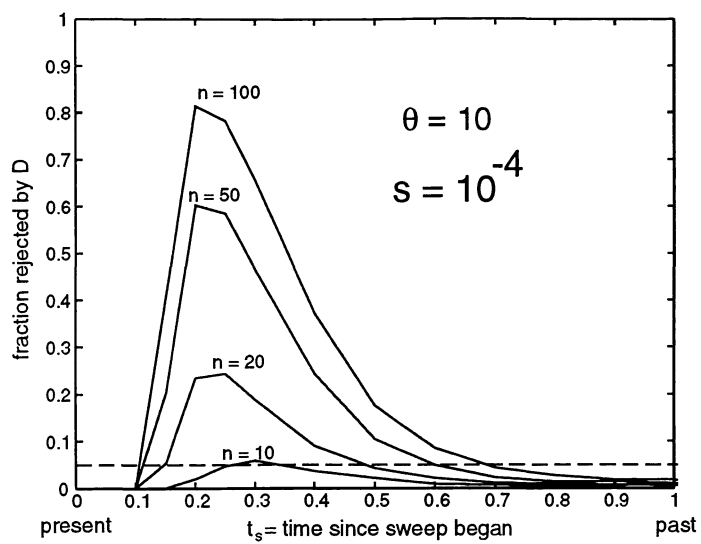
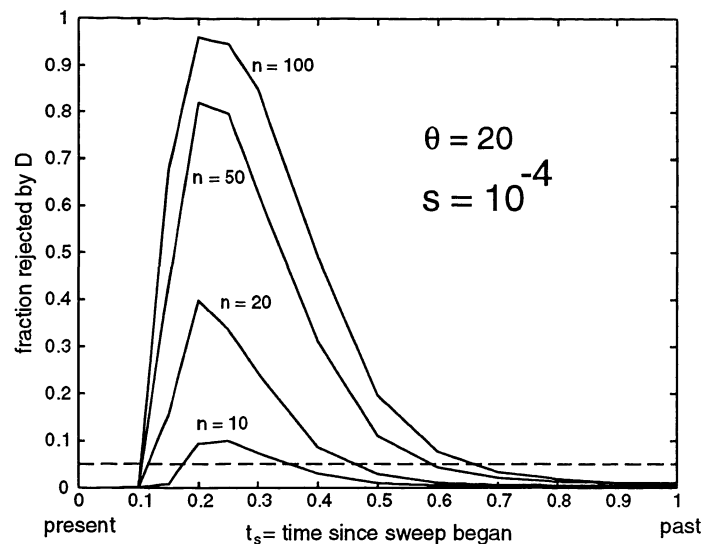


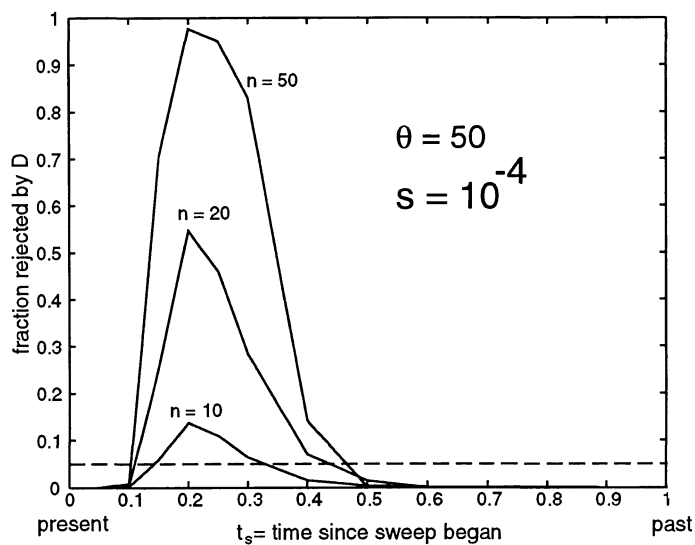
Figure 2



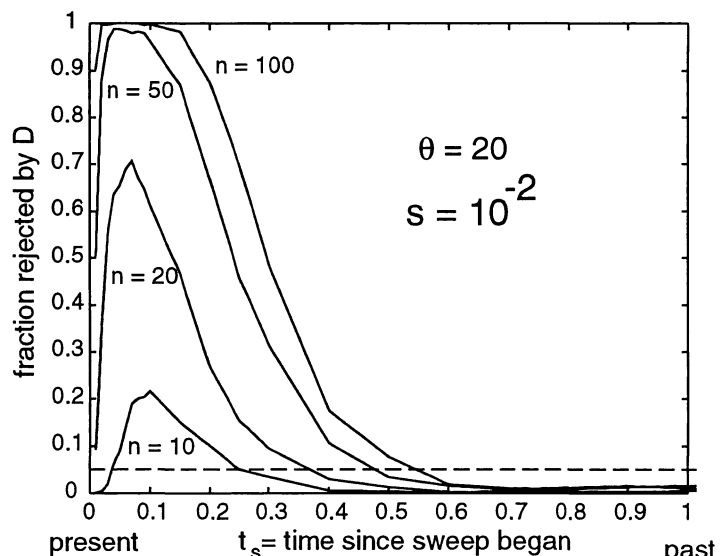
(a)



(b)

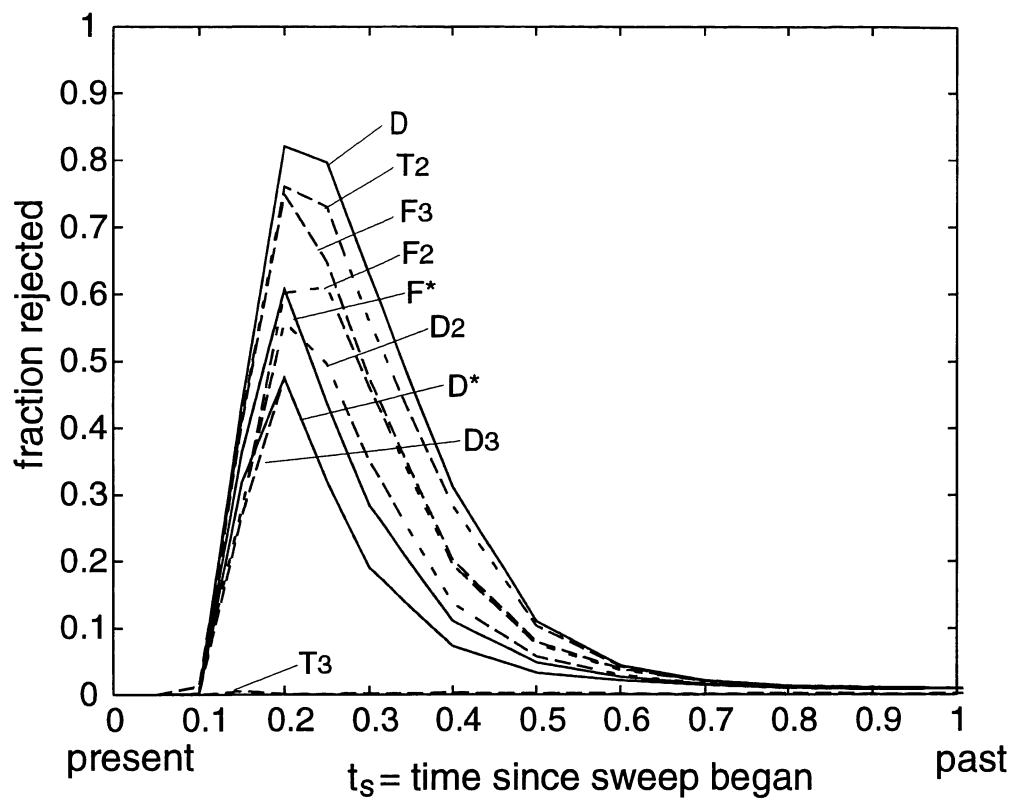


(c)

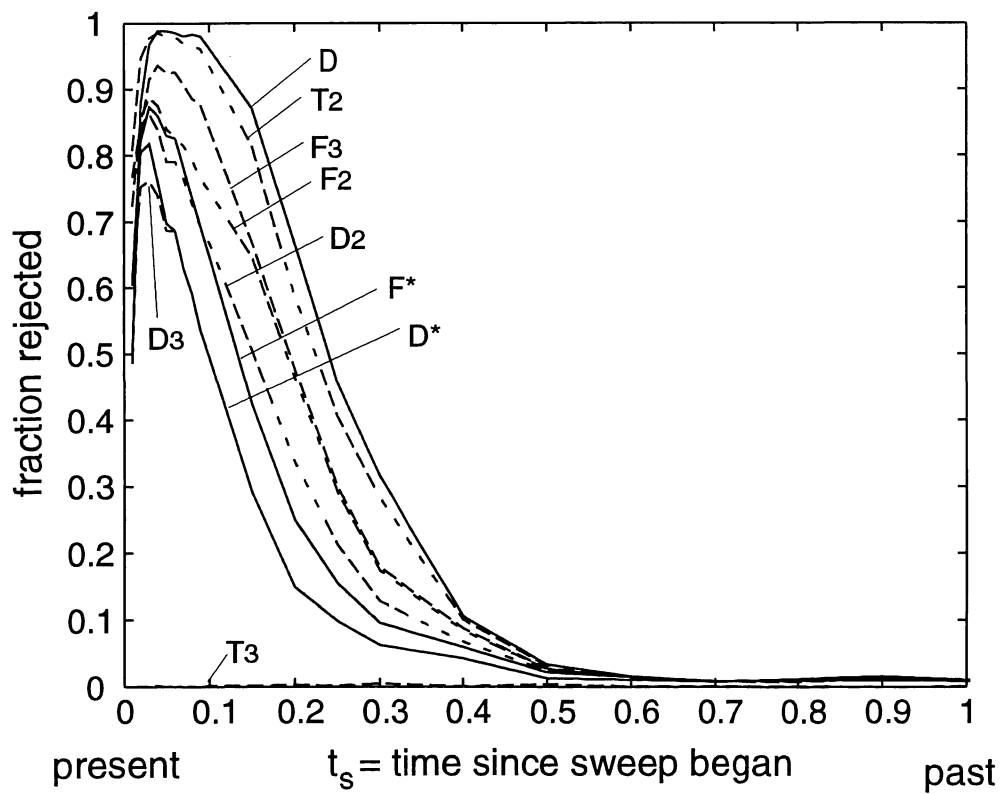


(d)

Figure 3



(a)



(b)

Figure 4

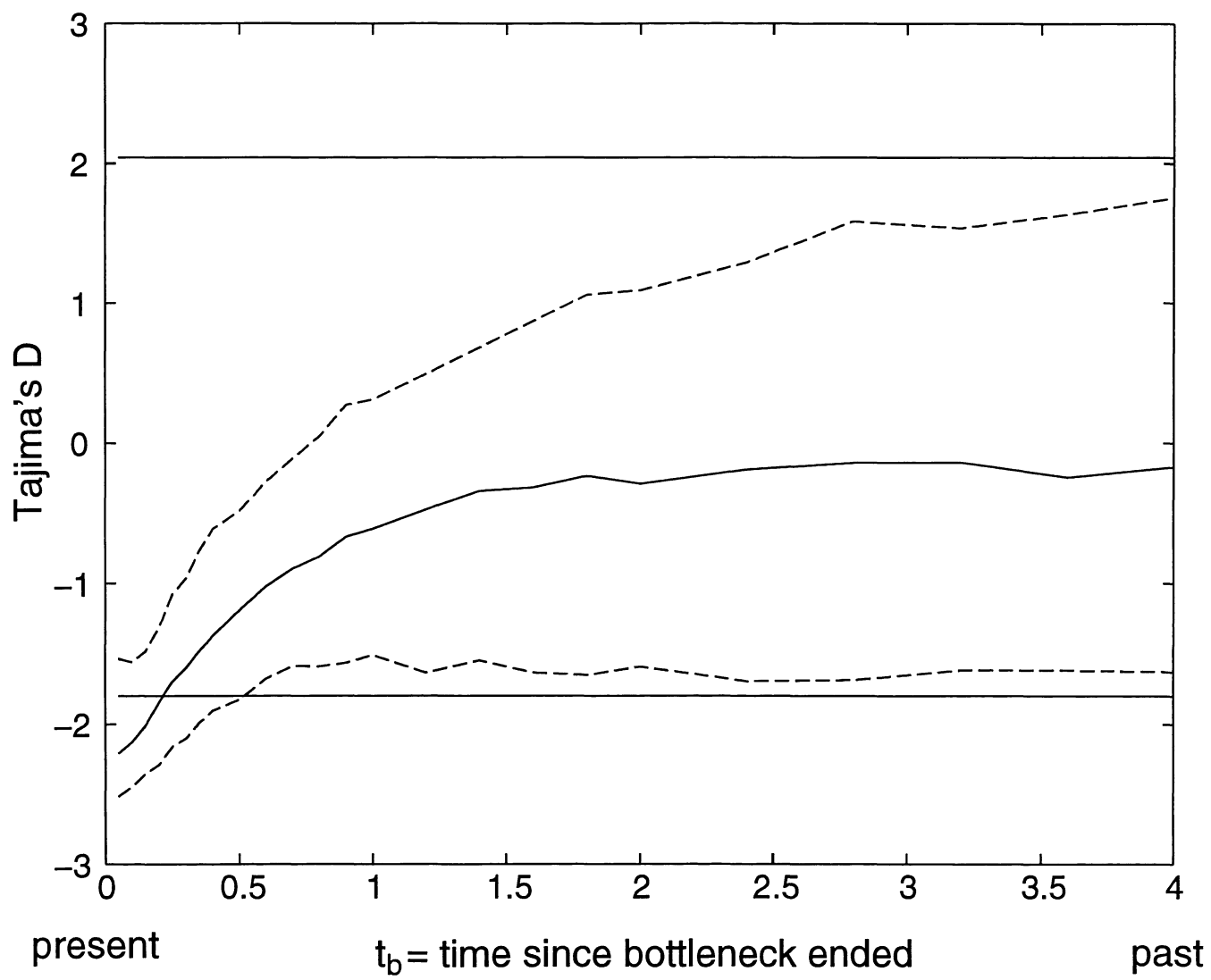
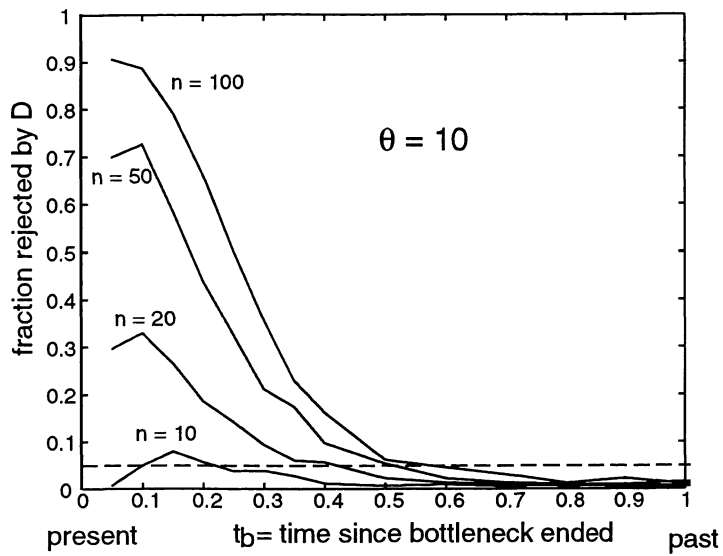
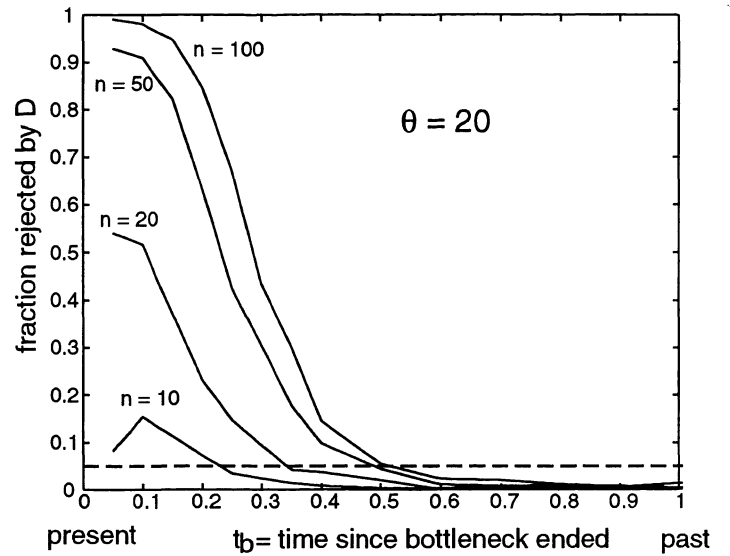


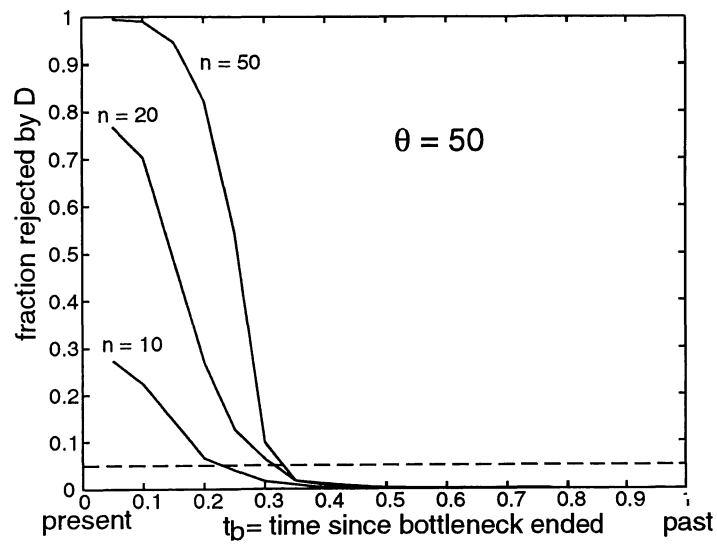
Figure 5



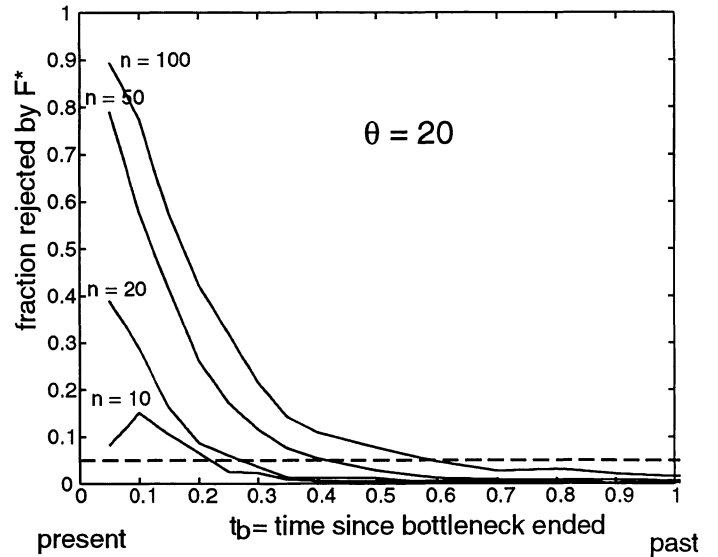
(a)



(b)



(c)



(d)

Figure 6

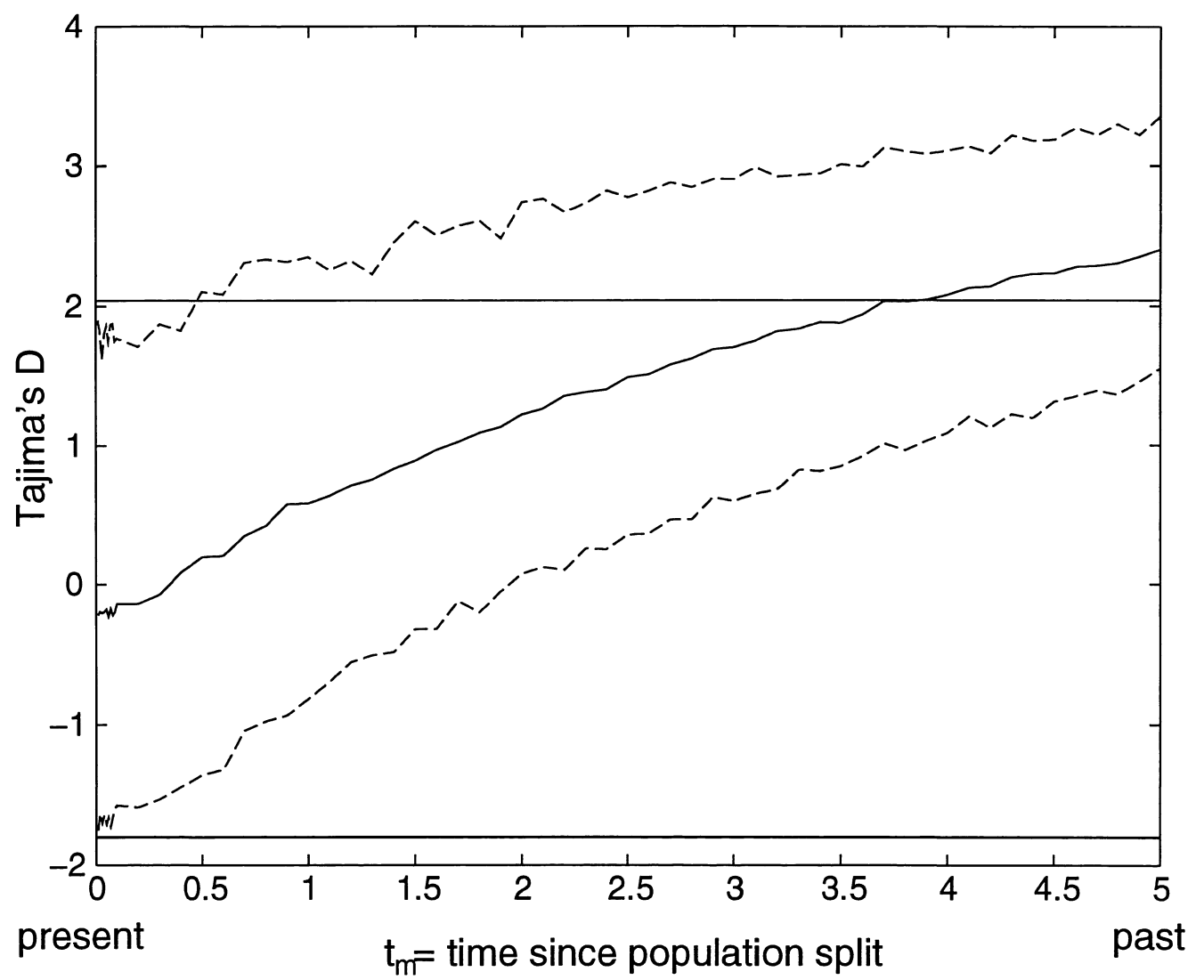


Figure 7

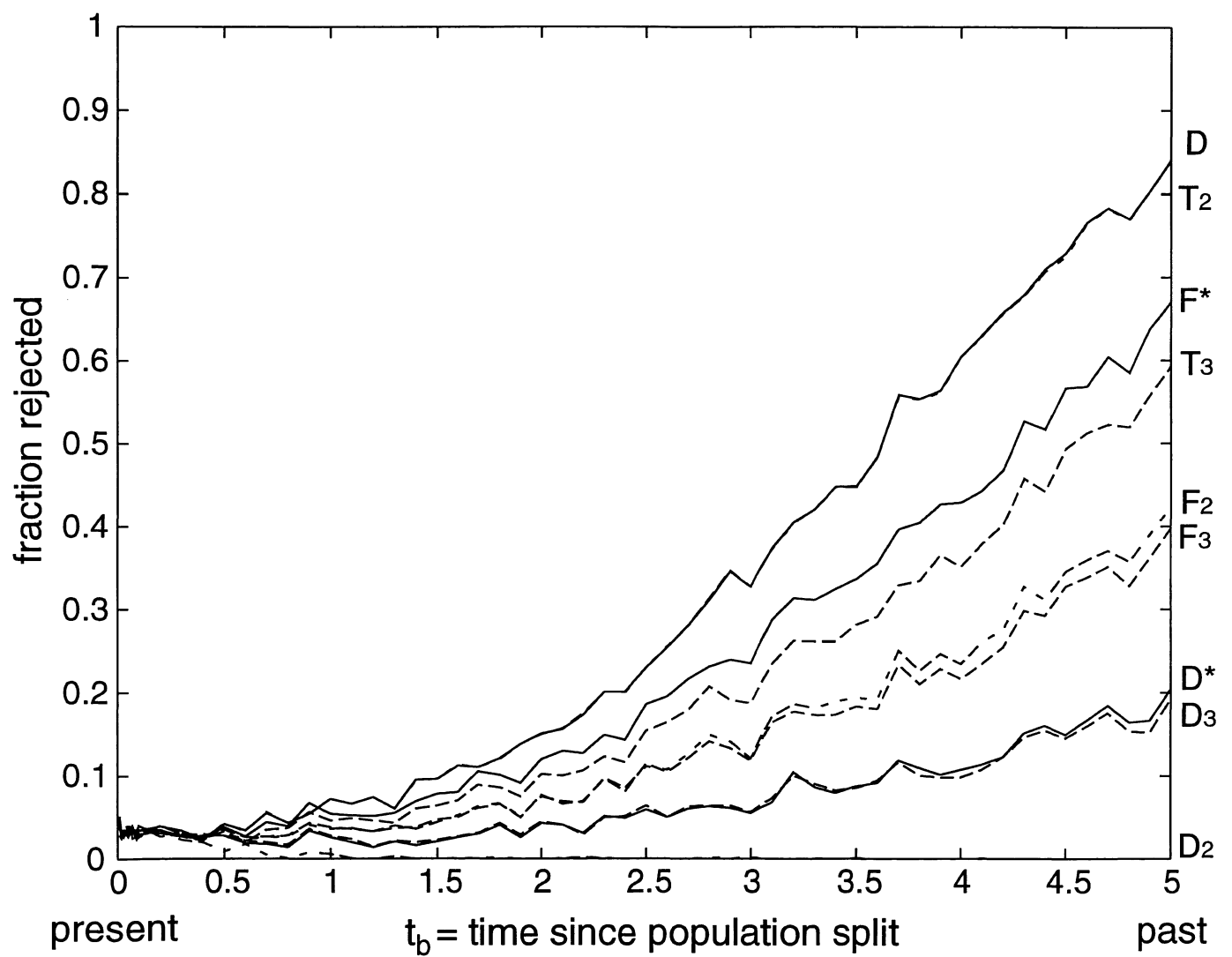


Figure 8

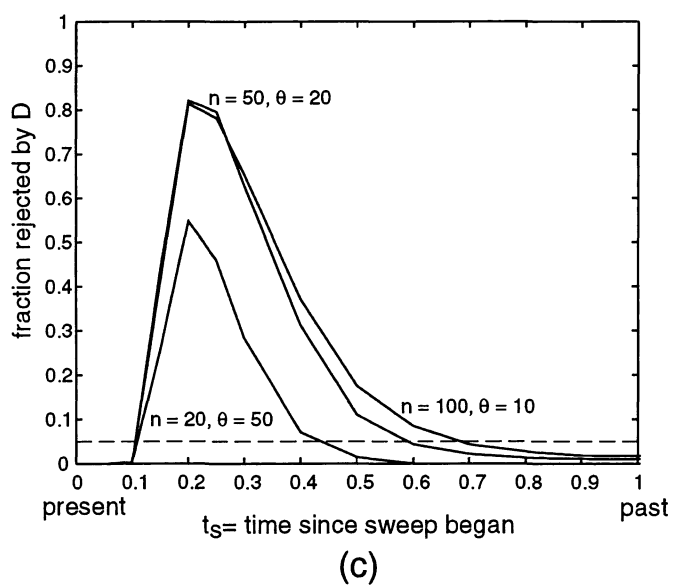
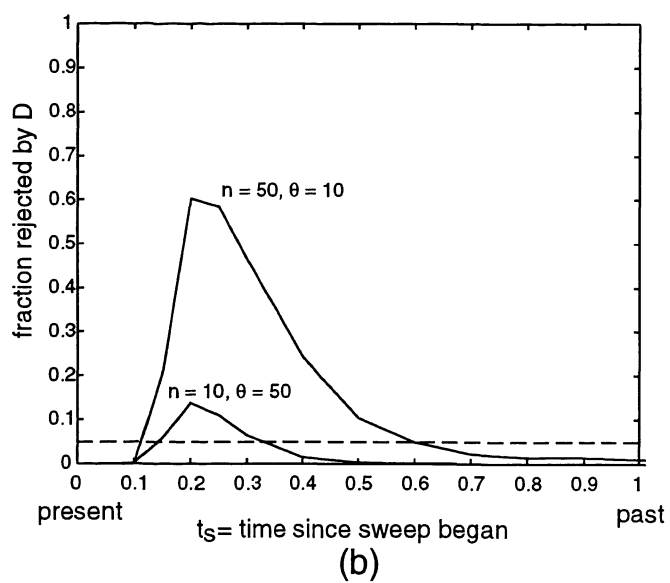
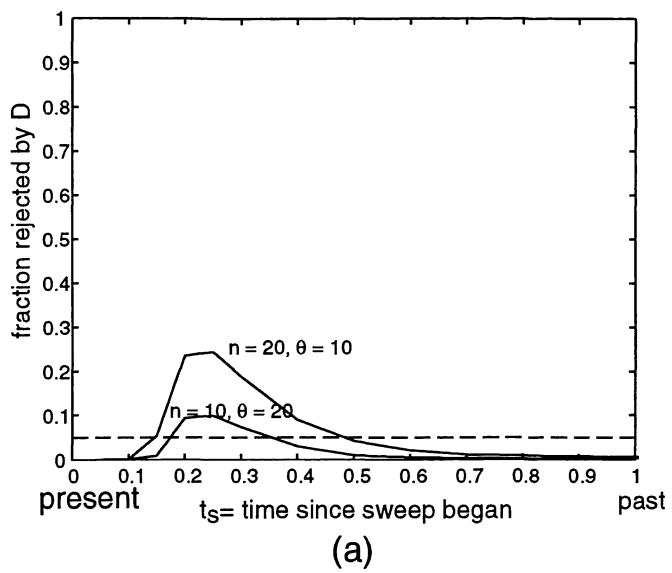


Figure 9

Table 1: Confidence intervals for θ given S when $n = 50$ and $\beta = 0.01$. $C\beta$ is a 99% CI for θ .

S	θ_L	θ_U	S	θ_L	θ_U	S	θ_L	θ_U	S	θ_L	θ_U	S	θ_L	θ_U
0	0.0	1.3	40	3.8	20.0	80	8.2	37.7	120	12.5	55.3	160	16.9	72.9
1	0.0	2.1	41	3.9	20.5	81	8.3	38.1	121	12.6	55.7	161	17.0	73.3
2	0.0	2.6	42	4.0	20.9	82	8.4	38.6	122	12.7	56.2	162	17.1	73.8
3	0.0	3.2	43	4.2	21.4	83	8.5	39.0	123	12.8	56.6	163	17.2	74.2
4	0.1	3.7	44	4.3	21.8	84	8.6	39.5	124	13.0	57.1	164	17.3	74.6
5	0.2	4.2	45	4.4	22.3	85	8.7	39.9	125	13.1	57.5	165	17.4	75.1
6	0.3	4.7	46	4.5	22.7	86	8.8	40.3	126	13.2	57.9	166	17.5	75.5
7	0.3	5.1	47	4.6	23.1	87	8.9	40.8	127	13.3	58.4	167	17.6	76.0
8	0.4	5.6	48	4.7	23.6	88	9.0	41.2	128	13.4	58.8	168	17.7	76.4
9	0.5	6.1	49	4.8	24.0	89	9.1	41.7	129	13.5	59.3	169	17.9	76.8
10	0.6	6.6	50	4.9	24.5	90	9.3	42.1	130	13.6	59.7	170	18.0	77.3
11	0.7	7.0	51	5.0	24.9	91	9.4	42.5	131	13.7	60.1	171	18.1	77.7
12	0.8	7.5	52	5.1	25.4	92	9.5	43.0	132	13.8	60.6	172	18.2	78.2
13	0.9	7.9	53	5.2	25.8	93	9.6	43.4	133	13.9	61.0	173	18.3	78.6
14	1.0	8.4	54	5.3	26.2	94	9.7	43.9	134	14.0	61.5	174	18.4	79.0
15	1.1	8.9	55	5.5	26.7	95	9.8	44.3	135	14.2	61.9	175	18.5	79.5
16	1.3	9.3	56	5.6	27.1	96	9.9	44.7	136	14.3	62.3	176	18.6	79.9
17	1.4	9.8	57	5.7	27.6	97	10.0	45.2	137	14.4	62.8	177	18.7	80.4
18	1.5	10.2	58	5.8	28.0	98	10.1	45.6	138	14.5	63.2	178	18.8	80.8
19	1.6	10.7	59	5.9	28.4	99	10.2	46.1	139	14.6	63.7	179	18.9	81.2
20	1.7	11.1	60	6.0	28.9	100	10.3	46.5	140	14.7	64.1	180	19.0	81.7
21	1.8	11.6	61	6.1	29.3	101	10.5	46.9	141	14.8	64.5	181	19.2	82.1
22	1.9	12.0	62	6.2	29.8	102	10.6	47.4	142	14.9	65.0	182	19.3	82.6
23	2.0	12.5	63	6.3	30.2	103	10.7	47.8	143	15.0	65.4	183	19.4	83.0
24	2.1	12.9	64	6.4	30.7	104	10.8	48.3	144	15.1	65.9	184	19.5	83.4
25	2.2	13.4	65	6.5	31.1	105	10.9	48.7	145	15.2	66.3	185	19.6	83.9
26	2.3	13.8	66	6.6	31.5	106	11.0	49.1	146	15.3	66.7	186	19.7	84.3
27	2.4	14.3	67	6.8	32.0	107	11.1	49.6	147	15.5	67.2	187	19.8	84.8
28	2.5	14.7	68	6.9	32.4	108	11.2	50.0	148	15.6	67.6	188	19.9	85.2
29	2.6	15.2	69	7.0	32.9	109	11.3	50.5	149	15.7	68.1	189	20.0	85.6
30	2.7	15.6	70	7.1	33.3	110	11.4	50.9	150	15.8	68.5	190	20.1	86.1
31	2.9	16.0	71	7.2	33.7	111	11.5	51.3	151	15.9	68.9	191	20.2	86.5
32	3.0	16.5	72	7.3	34.2	112	11.7	51.8	152	16.0	69.4	192	20.4	86.9
33	3.1	16.9	73	7.4	34.6	113	11.8	52.2	153	16.1	69.8	193	20.5	87.4
34	3.2	17.4	74	7.5	35.1	114	11.9	52.7	154	16.2	70.3	194	20.6	87.8
35	3.3	17.8	75	7.6	35.5	115	12.0	53.1	155	16.3	70.7	195	20.7	88.3
36	3.4	18.3	76	7.7	35.9	116	12.1	53.5	156	16.4	71.1	196	20.8	88.7
37	3.5	18.7	77	7.8	36.4	117	12.2	54.0	157	16.5	71.6	197	20.9	89.1
38	3.6	19.2	78	8.0	36.8	118	12.3	54.4	158	16.7	72.0	198	21.0	89.6
39	3.7	19.6	79	8.1	37.3	119	12.4	54.9	159	16.8	72.4	199	21.1	90.0

Table 2: Coefficients of Linear Approximations to a 1- β Confidence Interval for θ $C\beta = [bS + c, qS + r]$					
β	n	b	c	q	r
0.01	10	0.133	-0.484	1.236	3.787
	20	0.121	-0.481	0.709	2.858
	50	0.108	-0.474	0.441	2.302
	100	0.101	-0.484	0.341	2.039
0.001	10	0.102	-0.483	1.782	6.304
	20	0.094	-0.473	0.904	4.418
	50	0.087	-0.468	0.519	3.408
	100	0.081	-0.856	0.389	3.420

Table 3: Level 0.05 Critical Values of Tajima's D Test Based on a 99% Confidence Interval for θ given S

n = 10			n = 20			n = 50			n = 100		
S	DL	Du	S	DL	Du	S	DL	Du	S	DL	Du
0	-1.79	1.84	0	-1.78	1.97	0	-1.70	2.11	0	-1.58	2.21
1-26	-1.80	1.84	1-3	-1.82	1.97	1-22	-1.77	2.11	1-24	-1.70	2.21
27-41	-1.80	1.83	4-14	-1.83	1.97	23-31	-1.77	2.06	25-34	-1.70	2.15
42-48	-1.80	1.81	15-20	-1.84	1.97	32-41	-1.77	2.00	35-44	-1.70	2.07
49-63	-1.79	1.79	21-28	-1.84	1.96	42-50	-1.73	1.97	45-74	-1.70	2.04
64-71	-1.78	1.78	29-36	-1.84	1.90	51-73	-1.73	1.95	75-78	-1.68	2.01
72-135	-1.78	1.74	37-45	-1.84	1.88	74-155	-1.75	1.95	79-159	-1.69	2.01
			46-86	-1.84	1.87						
			87-144	-1.85	1.87						
			145-147	-1.85	1.82						
beta	-1.733	1.975	beta	-1.803	2.001	beta	-1.800	2.044	beta	-1.781	2.073

Table 4: Level 0.05 Critical Values of Fu and Li's D* Test Based on a 99% CI for θ given S

n = 10			n = 20			n = 50			n = 100		
S	D*L	D*U	S	D*L	D*U	S	D*L	D*U	s	D*L	D*U
0	-2.06	1.41	0	-2.4	1.40	0	-2.57	1.48	0	-2.68	1.32
1-48	-2.08	1.42	1-2	-2.49	1.44	1-13	-2.58	1.51	1	-2.68	1.51
49-63	-2.08	1.40	3-7	-2.59	1.44	14-17	-2.59	1.51	2-4	-2.68	1.55
64-78	-2.06	1.36	8-13	-2.67	1.44	18-19	-2.61	1.51	5-24	-2.68	1.59
79-861	-2.06	1.35	14-41	-2.70	1.44	20-24	-2.71	1.51	25-44	-2.54	1.59
87-108	-2.06	1.34	42-45	-2.73	1.44	25-42	-2.72	1.51	45-49	-2.50	1.59
109-135	-2.06	1.32	46-53	-2.73	1.43	43-50	-2.76	1.51	50-52	-2.52	1.59
			54-61	-2.73	1.42	51-60	-2.76	1.50	53-58	-2.54	1.59
			62	-2.73	1.38	61-68	-2.80	1.50	59-64	-2.56	1.59
			63-84	-2.76	1.38	69-71	-2.80	1.45	65-74	-2.56	1.57
			85-86	-2.78	1.38	72-73	-2.84	1.45	75-103	-2.56	1.54
			87-102	-2.78	1.36	74-77	-2.92	1.45	104-123	-2.56	1.51
			103-111	-2.78	1.35	78-114	-2.92	1.41	124-143	-2.56	1.49
			112-135	-2.78	1.34	115-151	-2.92	1.39	144-146	-2.57	1.49
			136-144	-2.78	1.33	152-155	-2.92	1.35	147-159	-2.58	1.49
(1993)	-2.02	1.38	(1993)	-2.43	1.37	(1993)	-2.45	1.44	(1993)	-2.33	1.53

Table 5: Level 0.05 Critical Values of F_U and L_i 's F^* Test Based on a 99% CI for θ given S

n = 10			n = 20			n = 50			n = 100		
S	F^*_L	F^*_U	S	F^*_L	F^*_U	S	F^*_L	F^*_U	s	F^*_L	F^*_U
0	-2.22	1.60	0	-2.54	1.65	0	-2.57	1.74	0	-2.52	1.67
1-48	-2.26	1.61	1-2	-2.62	1.67	1-6	-2.60	1.74	1-24	-2.52	1.83
49-63	-2.26	1.58	3-7	-2.69	1.67	7-19	-2.61	1.74	25-44	-2.47	1.83
64-71	-2.25	1.57	8-13	-2.74	1.67	20-24	-2.62	1.74	45-64	-2.42	1.83
72-101	-2.25	1.53	14-42	-2.76	1.67	25-41	-2.72	1.74	65-103	-2.40	1.83
102-108	-2.25	1.52	43-45	-2.78	1.67	42-50	-2.72	1.72	104-113	-2.40	1.82
109-135	-2.25	1.51	46-61	-2.78	1.62	51-60	-2.72	1.71	114-125	-2.40	1.81
			62	-2.78	1.58	61-68	-2.77	1.71	126-159	-2.43	1.81
			63-102	-2.81	1.58	69-73	-2.77	1.70			
			103-144	-2.81	1.56	74-133	-2.85	1.70			
			145-147	-2.81	1.55	134-155	-2.85	1.68			
(1993)	-2.21	1.59	(1993)	-2.57	1.61	(1993)	-2.43	1.66	(1993)	-2.30	1.73